DOCUMENT RESUME

ED 385 569 TM 024 007

AUTHOR Hale, Gordon A.
TITLE Effects of Amount of Time Allowed on the Test of Written English.
INSTITUTION Educational Testing Service, Princeton, N.J.
REPORT NO ETS-RR-92-27; TOEFL-RR-39
PUB DATE Jun 92
NOTE 50p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *College Students; English; *Essay Tests; Higher Education; Language Proficiency; Limited English Speaking; Scores; *Student Attitudes; Testing; *Test Results; *Timed Tests; Time Management
IDENTIFIERS *Test of Written English

ABSTRACT

This study examined students' essay performance on topics from the Test of Written English (TWE) under time limits of 30 minutes, as currently administered, and 45 minutes. In the main groups of the study, each student wrote an essay on one topic under the current time limit and on another under the 45-minute time limit. A total of 820 intensive English and academic international students participated. The correlation between scores for the time conditions was relatively high and approached the parallel-form reliability of the task, as determined by data from students who wrote essays on different topics under the same time limits. The provision of additional time apparently had little effect on the standings of the students in relation to each other. Mean scores on the TWE were about one-fourth to one-third point higher for the 45-minute condition, indicating a modest but reliable increase in scores. The magnitude of the effect was roughly comparable for students of low and high ability. Students regarded 45 minutes as more sufficient for accomplishing the task than 30 minutes. Practical implications are discussed. Appendixes contain the topics and the scoring guide. Seven tables present study findings. (Contains 12 references.)
(Author/SLD)

TEST OF ENGLISH AS A FOREIGN LANGUAGE

Research Reports

REPORT 39
June 1992

Effects of Amount of Time Allowed on the Test of Written English

Gordon Hale

EDUCATIONAL TESTING SERVICE

Effects of Amount of Time Allowed on the Test of Written English

Gordon A. Hale

Educational Testing Service
Princeton, New Jersey

RR-92-27

Educational Testing Service is an Equal Opportunity/Affirmative Action Employer.

The Test of English as a Foreign Language (TOEFL®) was developed in 1963 by a National Council on the Testing of English as a Foreign Language, which was formed through the cooperative effort of more than thirty organizations, public and private, that were concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program, and in 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations (GRE) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education.

ETS administers the TOEFL program under the general direction of a Policy Council that was established by, and is affiliated with, the sponsoring organizations. Members of the Policy Council represent the College Board and the GRE Board and such institutions and agencies as graduate schools of business, junior and community colleges, nonprofit educational exchange agencies, and agencies of the United States government.

A continuing program of research related to the TOEFL test is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the Policy Council, the TOEFL Committee of Examiners, and distinguished English as a second language specialists from the academic community. Currently the Committee meets twice yearly to review and approve proposals for test-related research and to set guidelines for the entire scope of the TOEFL research program. Members of the Research Committee serve three-year terms at the invitation of the Policy Council; the chair of the committee serves on the Policy Council.

Because the studies are specific to the test and the testing program, most of the actual research is conducted by ETS staff rather than by outside researchers. However, many projects require the cooperation of other institutions, particularly those with programs in the teaching of English as a foreign or second language. Representatives of such programs who are interested in participating in or conducting TOEFL-related research are invited to contact the TOEFL program office. All TOEFL research projects must undergo appropriate ETS review to ascertain that the confidentiality of data will be protected.

Current (1991-92) members of the TOEFL Research Committee are:

James Dean Brown	University of Hawaii
Patricia Dunkel (Chair)	Pennsylvania State University
William Grabe	Northern Arizona University
Kyle Perkins	Southern Illinois University at Carbondale
Elizabeth C. Traugott	Stanford University
John Upshur	Concordia University

Acknowledgments

The author wishes to express his sincere appreciation to:

Brent Bridgeman and Carol Taylor for helpful suggestions in planning and conducting the study,
Donald Rock for statistical advice,
Wing Lowe and Bruce Kaplan for programming the data analyses,
Karin Steinhaus for help in selecting the test items,
Pamela Bowski, Joyce Gant, and Chris Taylor for assistance in preparing for data collection and tabulating the data,
Eleanore DeYoung and Joanne Farr for secretarial assistance, and
Roberta Camp, Eduardo Cascallar, Gerald DeMauro, Grant Henning, and Linda Tang for reviewing earlier drafts of this report.

The author also gratefully acknowledges the assistance of the following persons for their invaluable roles in the essay reading:

Nancy Olsen for organizing the reading,
Mary Bly, Tom Lederer and Julie Pratico for managing the reading
Bonnie Katz and Barbara Voltmer for providing advice and assistance in preparation for the reading, and
The 18 readers for their untiring efforts throughout the reading.

Special thanks are extended to the on-site project coordinators, listed below, who contributed generously of their time in arranging and carrying out the data-collection activities.

Gaye Childress and Rebecca Smith, Intensive English Language Institute, University of North Texas
Gilbert Couts, English Language Institute, American University
Marc Cummings, Intensive English as a Second Language Program, University of Louisville
Larry Francis, Intensive English Program, University of Missouri, Columbia
Linda Gould, Intensive English Institute, University of Illinois
Gail Kellersberger, English Language Institute, University of Houston, Downtown
James Stalker and Ralph Barrett, English Language Center, Michigan State University
Anne Wyatt-Brown, English Language Institute, University of Florida
Paul Angelis, Southern Illinois University
Louis Arena, University of Delaware
Millie Audas and Jane Hughey, University of Oklahoma
Emily Catherine Day and Paul Webb, Eastern Michigan University
Dennis Evans, Diane Clymer and Boyer Rickel, University of Arizona
Robert Kantor, The Ohio State University

Finally, the author is indebted to the TOEFL Research Committee for its support of the project and for constructive comments on an earlier draft of this report.

Abstract

This study examined students' essay performance on topics from the Test of Written English (TWE®) under two time limits--30 minutes, as on the current TWE, and 45 minutes. In the main groups of the study, each student wrote an essay on one topic under one time limit and on another topic under the other time limit (with orders counterbalanced). The correlation between scores for 30- versus 45-minute conditions was relatively high and approached the parallel-form reliability of the task, as indicated by data from students who wrote essays on separate topics under the same time limits. Thus, the provision of additional time apparently had little effect on the standing of the students in relation to each other. Both parallel-form reliability and interrater reliability were approximately the same for the 30- and 45-minute conditions. Mean scores on the 6-point TWE scale were found to be higher by about 1/4 to 1/3 point under the 45-minute condition than the 30-minute condition, indicating that provision of additional time produced a modest but reliable increase in scores. The magnitude of the effect was roughly comparable for students of low versus high proficiency, and for students in intensive English programs versus students in academic coursework. Responses to a questionnaire indicated that the students regarded 45 minutes as more sufficient for accomplishing the task than 30 minutes. The results are discussed in relation to the literature on time effects and to practical implications for the Test of Written English.

Table of Contents

List of Tables

Introduction

The Test of Written English (TWE) is an essay test for nonnative English speakers that was developed to accompany the Test of English as a Foreign Language (TOEFL®) at selected administrations. Its purpose is to assess students' ability to communicate in written English, within reasonable limits allowable in a standardized testing situation. Students are permitted 30 minutes for completion of their essays.

Although the TWE is effective as currently administered, continued research on the test and its conditions of administration can help ensure that the test is as valid as possible. In an agenda for further research on the TWE, Stansfield and Ross (1988) proposed that the effects of extending the time limit be examined, in order to obtain more empirical evidence as to the relative merits of the 30-minute time allotment. Toward that end, the present study examined the performance of nonnative English speakers on TWE topics under a 45-minute limit as well as the current 30-minute limit. More specifically, students in the main treatment groups were asked to write essays on two topics, one in a 30-minute condition, and the other in a 45-minute condition (with topics counterbalanced across groups).

Several issues were of interest. A primary issue concerned the correlation in performance under the two time conditions. If the correlation were found to be low in relation to reliability, it could be inferred that allotment of an additional 15 minutes beyond the current 30 minutes altered the ranking of students in relation to each other and, to that extent, changed the character of the test. On the other hand, if the correlation were high in relation to reliability, the students' relative standing on the test would not have been markedly affected by the provision of extra time, and in that respect, the measurement character of the test would not have been changed. To assess the parallel-form reliability of the test, and thus provide a basis for comparison with the correlation between time conditions, additional groups were asked to write essays on the two topics under the same time limit for both topics--a 30-minute time limit in both cases for one group, and a 45-minute time limit in both cases for another group.

A second issue concerned the comparison between reliabilities observed under the two time limits. In this regard, the information on parallel-form reliability was useful, not only as an aid in interpreting the correlation between time conditions, but as valuable evidence in its own right. A substantial reliability difference in favor of the 45-minute condition would be evidence that this condition provides greater consistency of measurement than does the 30-minute time condition. Also, data on interrater reliabilities would contribute information related to consistency in the scoring process and possible differences therein for essays written under the two time limits.

A third issue was whether students' essay scores would be changed if the students were given the extra 15 minutes and, if so, by how much. If scores were found to be higher under the 45- than the 30-minute time limit, one possible interpretation might be that the longer time limit allows a fuller opportunity for students to demonstrate their written communication skills. It should be noted that, if the TWE were a norm-referenced test,

data regarding mean performance would be relatively insignificant in relation to the correlational data in addressing the question of whether extra time affects the psychometric integrity of the test. For example, even if mean scores were increased with additional time, evidence that the relative standing of students was unaffected by extra time would suggest that the essential measurement properties of the test remained unchanged. However, the TWE scale is criterion referenced, in that each level of the 6-point TWE scale is associated with a different set of descriptors regarding the student's writing competence (see Test of Written English Scoring Guide in Appendix B). Because TWE scores can therefore be meaningfully interpreted in absolute as well as relative terms, it is important to determine whether the mean score received, as well as the relative standing of students, is affected by the provision of additional time.

A fourth issue of interest involved the students' reactions to the 30- and 45-minute time limits. If the students found 45 minutes to be more satisfactory than 30 minutes, it could be argued that the provision of additional time increases the test's face validity from the students' standpoint. To address this issue, students responded to a questionnaire (after completing their essays) in which they were asked about the adequacy of the 30- and 45-minute time limits. They were also asked how they used the additional time provided in the 45-minute condition.

Although the primary effects of interest were those involving students across the entire range of proficiency, for reasons discussed below under "Relevant Literature" it was also of interest to determine whether the time allotment would have different effects on mean scores for students of low versus high proficiency. Thus, analyses looked for differential effects for students scoring low versus high on the present essays. Analyses also examined differential effects for students taking regular academic coursework versus students enrolled in intensive English programs--that is, programs of instruction in English for students who are not yet proficient enough in the language to qualify for matriculation into a regular academic program. The distinction between these groups serves as a rough, independent index of English proficiency.

In studying the effects of time limits, samples of each of the two general categories of essay topics that have been used in the TWE were employed here. These are termed "prose" topics and "chart/graph" topics.[1] In recent years the TWE has employed prose topics exclusively, so that the results involving this topic type are of primary interest in this study. Nevertheless, it is of value to examine the results for chart/graph as well

[1]When the TWE was first developed, the two topic categories used were "compare/contrast, take-a-position" and "chart/graph"; the former involved comparison of two possible points of view, with argument in favor of one; the latter involved discussion based on the data in a graph or chart. More recently, the first category has been expanded to include a variety of formats and, hence, are more appropriately labeled "prose" topics, to differentiate them from topics in which the stimulus includes graphic material in addition to text.

as prose topics, to determine the extent to which the effects observed here are consistent across topics that differ in format.

Relevant Literature

The issue of the correlation between time conditions and its relation to reliability apparently has not been addressed in research conducted to date. Nor has research been done that compared time conditions with respect to reliability or systematically assessed students' perceptions of the adequacy of the time limits. Research varying the time limits of essay tests has focused principally on the effects of the time limits on mean performance.

Most of the research on time effects has been conducted with English speakers, although a few studies have included nonnative English speakers. Some studies have observed a significant effect of the time limit on essay performance. Biola (1982) found that college freshmen achieved better scores under a 120-minute than a 45-minute time limit. Younkin (1986), in a study involving native and nonnative English-speaking college students, found that a combination of groups given 20 or 10 extra minutes received higher scores than those given a baseline time estimated to be approximately 50 minutes. (The baseline time for the essay was not indicated but has been inferred by the present author from the overall time specified for the essay plus other subtests administered with it.) The effects in both of these studies were reported to be significant, although the magnitudes of the effects in relation to the score scale were not given. In the study by Younkin, the benefits of extra time were found to be roughly equal for native and nonnative English speakers.

Other studies have found no significant effect of the time limit, although in some cases there was a tendency toward an effect. Livingston (1987) found that high school and college students given 30 minutes to write an essay did not score significantly higher than those given 20 minutes, although there was a trend toward an effect (of about 1/3 to 1/2 point on a 12-point scale) for higher proficiency students. Two studies with nonnative English speakers (Caudery, 1990; Kroll, 1990) compared essays written in class versus those written at home; both studies used relatively small sample sizes. Kroll examined 25 students' scores on (a) two essays written in 60 minutes versus (b) two essays written at home over a 10-14 day period. The overall difference (.4 points on a 6-point scale) was not significant, although a consistent score difference in favor of the longer time period was observed for each of the five language groups represented. Caudery, studying 24 adolescent students in Cyprus, found a nonsignificant difference (.24 points on a 20-point scale) between scores on English compositions when the students were allowed 40 minutes (in class) versus two days (partly in class and partly at home). In a study of medical school applicants, Mitchell and Anderson (1987) reported that, in the opinion of essay scorers, provision of 45 minutes allowed for fuller essay development than did 30 minutes, although mean scores and significance tests were not reported.

A reasonable general conclusion is that, under certain conditions at least, the amount of time allowed has an effect on mean essay scores. And, apparently, an effect of the time limit can be expected for nonnative as well as for native English speakers. In fact, despite the above-mentioned study that obtained comparable effects for these two types of students, conceptually it seems reasonable to assume that nonnative English speakers would be especially susceptible to variation in time limits. As Kroll (1990) indicates, nonnative English speakers have difficulty with the code of English, as discussed by such writers as Collins and Gentner (1980) and McLaughlin (1987), so that extending the time allowed might be particularly beneficial for them.

Drawing on this last observation, one might hypothesize that students' proficiency, in general, may play a role in determining the effects of time limits. Low-proficiency writers may stand to gain the most from additional time. Assuming that these students take longer than high-proficiency writers to organize and express their thoughts, they may feel rushed with the shorter time limit. Provision of a more liberal time limit may, therefore, be especially beneficial for low-proficiency writers. On the other hand, the literature also provides a basis for an opposing hypothesis. In Livingston's (1987) study there was a tendency toward an effect for high-proficiency but not low-proficiency writers. Also, in a survey of students taking a 30-minute writing test, the proficient writers, more than the nonproficient writers, indicated a need for more time for planning, writing, and proofreading (Ruth & Murphy, 1988). Among the objectives of the present study was to determine if proficiency plays a role in determining effects of time limits and, if so, whether the greater benefit accrues to the low- or to the high-proficiency students.

Method

Subjects

A total of 820 international students were tested, 482 students in eight university-based intensive English programs and 338 students, both graduate and undergraduate, enrolled in academic coursework in six universities. Students from these two sources are herein referred to as "intensive English students" and "academic students."

The academic students were invited to participate via posted notices and newspaper advertisements asking for nonnative English-speaking international students. They were offered $20 as an incentive for participation. The intensive English students participated as a program activity and were each given a TOEFL Test Kit. Students from both sources were included in order to produce a sample with a wide range of English proficiency. Academic students have standardized test scores (usually on TOEFL) that are high enough to permit entry into regular undergraduate or graduate coursework. Intensive English students, on the other hand, are generally not proficient enough in English to undertake academic coursework, at least on a full-time basis.

A total of 62% of the students were college graduates. All were nonnative English speakers. Native languages of the students were Chinese dialects (24% of the sample), Japanese (20%), Spanish (13%), Korean (7%), Arabic (6%), Thai (4%), and 54 other language groups with fewer than 4% in each. Native countries represented were Japan (19% of the sample), the People's Republic of China (11%), Taiwan (9%), Republic of Korea (7%), India (5%), Thailand (4%), and 80 other countries with fewer than 4% from each. TOEFL and Michigan Test scores were requested, but not enough students had scores to permit meaningful analysis of them. (The scores provided were generally for academic students.)

Materials

Essays. Special four-page sealed booklets were prepared for writing the essays. The front page contained the printed instructions that are distributed to students when they take the Test of Written English. The essay topic appeared at the top of the second page, followed by 2-1/2 pages of lines for writing the essays; the lines were of the same size and spacing as those on a TWE answer sheet but, to accommodate the extended time period, the booklet contained 50% more lines than the TWE answer sheet.

Four essay topics were used in the study, two "prose" topics and two "chart/graph" topics; in the former case, the stimulus or prompt about which the student was to write was presented in prose form, whereas in the latter case, the stimulus was a chart or graph, with an accompanying question to be addressed in the essay. The topics had been used previously in connection with operational TOEFL administrations some years before but had not been

disclosed or published in any form and, thus, were highly unlikely to have been seen by these students. The four topics are presented in Appendix A.

Questionnaire. In the questionnaire the students were asked for their reactions to the essay writing experience. The principal questions can be seen in Tables 5, 6, and 7 in the Results section, along with data indicating the students' responses. (The students were also asked about performance on the first versus second essay written, to aid interpretation in case an order effect were found. However, performance on the first and second essays did not differ, as indicated below, so that these questionnaire responses were essentially moot and are not presented here.)

Research Design

The treatment groups in the study are shown in Table 1; entries in the table are the numbers of students per subgroup. As shown in the top headings in Table 1, some students were given prose topics and others, chart/graph topics. Within those general groups, the topic order was counterbalanced such that some students wrote their first essay on one topic and their second essay on the other topic, while the reverse was true for other students.

Main treatment groups. In the main treatment groups (first two rows of Table 1) each student wrote on one essay topic under one time condition then wrote on another essay topic under the other time condition. The time order was counterbalanced, such that students in some groups were given a 30-minute time limit for the first essay and a 45-minute limit for the second, and students in other groups were given the time conditions in the reverse order. Note that, in an ideal design, the numbers per group would be equal, so that the different orders would completely balance each other in examining mean performance of the different groups. However, use of least squares analyses of variance effectively ensured that the different orders contributed equally in computing effects of the factors under study.

Groups included to assess parallel-form reliability. The other key treatment groups were those in which students wrote essays on two topics, each under the same time limit, to assess parallel-form reliability. Within each topic order, students were asked to write two 30-minute essays or two 45-minute essays.

Supplemental condition. A supplemental condition was also included to obtain pilot data on the role of forced planning. In this condition, the students were given 45 minutes for the first essay, with instructions to spend 15 minutes planning and 30 minutes writing. Performance on this "special" 45-minute essay was to be compared with performance on the regular 45-minute essay by students in the second row of Table 1. A second essay written under the standard 30-minute time limit served as a control, as discussed below in the Results section. This type of comparison has severe limitations, but the total sample available was not large enough to permit

Table 1

Research Design

	Prose Topics		Chart/Graph Topics	
	Topic A then B	Topic B then A	Topic C then D	Topic D then C
Main Treatment Groups				
30 min. essay then 45-min. essay	N = 33	N = 31	N = 33	N = 27
45-min. essay then 30-min. essay	N = 41	N = 38	N = 35	N = 37
Groups Included to Assess Parallel-Form Reliability				
Two 30-min. essays	N = 52	N = 47	N = 48	N = 46
Two 45-min. essays	N = 61	N = 59	N = 59	N = 56
Supplemental Condition				
Special 45-min. essay then 30-min. essay	N = 33	N = 31	N = 26	N = 27

the most effective test of the role of forced planning and also to permit comprehensive assessment of the study's central issues. At the least, however, this condition provided preliminary information, as a step toward more comprehensive research on the topic.

For each of the five rows shown in Table 1, the sample included both intensive English students and academic students. Students were drawn from eight intensive English programs, two for each row in the table (except only one program in the third and fifth rows), and from six academic programs, one for each row (except two in the fourth row).

Within each row in Table 1 students were assigned to the four subgroups by spiralling of test materials. That is, four sets of materials were prepared for distribution to students within the testing session at each site: (a) Topic A followed by Topic B, (b) Topic B followed by Topic A, (c) Topic C followed by Topic D, and (d) Topic D followed by Topic C. Exceptions were two intensive English programs at which students were split into two testing sessions; to prevent effects of communication among students in each such case, the prose topics (A and B) were administered in one session and the chart/graph topics (C and D) in the other.

Procedure

The students were told that they would be asked to write two essays on topics like those in the Test of Written English. They were also told that they would complete a questionnaire asking for their opinions about the essay writing experience. The students were informed that, while they would not receive official scores, it was important that they do their best, as this study would help the researchers improve the essay test.

After the students answered a few background questions, testing on the first essay began with a statement by the test supervisor about the amount of time allowed. The test instructions closely paralleled those of the Test of Written English. After a break of a few minutes, testing on the second essay began with a statement about the amount of time allowed, followed by a repetition of the essential test instructions.

When the essay testing was completed, the students were asked to complete the questionnaire. The students were dismissed when all had been given an opportunity to complete the questionnaire.

Essay Scoring

A special reading was conducted under the direction of the essay reading staff of the Educational Testing Service Bay Area Office. The procedures and scoring guidelines used in scoring of operational forms of the Test of Written English were employed (cf., Educational Testing Service, 1989.) There were two groups of nine readers each, with a leader for each group, as well as a chief reader, all chosen from among experienced readers of TWE essays. The chart/graph essays were read first, with Topic C

assigned to one group and Topic D to the other. Then the prose essays were read, with Topic A assigned to one group and Topic B to the other. Among the advantages of assigning different topics to different groups was that a given reader would not encounter both essays written by any given student.

As in operational TWE readings, each essay was scored holistically by two readers on a 6-point scale. Where the two readers assigned scores differing by one point, the student's score was the average of the two readers' scores. Assignment of scores differing by more than one point was termed a discrepancy, and, in such a case, the student's score was that assigned by a third reader, who was either the group leader or the chief reader. The discrepancy rate was the percentage of papers for which discrepant ratings were obtained.

Scoring of each essay was "blind" with respect to all important factors: (a) the time limit under which the essay was written, (b) the experimental condition to which a student was assigned, (c) the student's educational status, and (d) whether the essay was the first or second one written by the student. To ensure blind scoring, all identifying information on the test booklets was covered with opaque removable tape during the essay reading. For each topic all essay booklets had been shuffled before scoring.

Results

Interrater Reliability

Before considering the results of the principal analyses, it is useful to examine data relating to interrater reliability. As in operational administrations of the TWE, three statistics were computed: (a) the discrepancy rate, which is the percentage of papers for which the two initial ratings differed by two or more points, (b) the correlation between scores given by the two readers, and (c) the coefficient alpha reliability, which is computed as:

$$\text{Alpha} = 2(1 - [S1^2 + S2^2]/ST^2),$$

where $S1^2$ and $S2^2$ refer to the rating variances of the first and second readers, respectively, and ST^2 refers to the variance of the sum of the ratings.

The results are shown in Table 2. Data are first presented, within prose and chart/graph topic types, for each individual topic and for topics combined. Then, most pertinent to the central issue of the study, the data are presented for 30-minute and 45-minute conditions to show whether interrater reliability differed according to the time allotment. Included in the latter analyses were data for 30-minute and 45-minute essays for all groups except the supplemental condition.

It should be noted that all principal analyses in this study have been conducted separately for the prose and chart/graph topic types, rather than for the two topic types combined. This was done because of the exclusive use of prose topics in recent TWE administrations and the need to determine how extra time affects performance in this topic type in particular; data for the chart/graph topic help establish the generalizability across topic types of the effects observed.

The discrepancy rates were somewhat higher, and the interrater correlations and alpha reliabilities somewhat lower, than typically observed with an operational TWE. (For the four TWE administrations during testing year 1990-91, averages for these three statistics were .01, .82, and .90, respectively.) This may have been due to the relatively small number of essay readers used in this study. With only 10 readers per topic, the interrater statistics can be much more noticeably affected by a reader who is generally lenient (or strict) than would be the case in a typical TWE reading, in which there are several times as many readers as in the present study. The effect of the small number of readers is especially apparent in the case of chart/graph Topic C. According to the supervisor of the essay reading, the higher discrepancy rate for Topic C than for the others (and, thus, lower interrater correlation and alpha reliability) was likely due to the fact that, by chance, readers who tended to assign low scores happened to be paired with readers who tended to assign high scores for several

Table 2

Discrepancy Rate, Interrater Correlation and Alpha Reliability for Major Groups of Papers

	Discrepancy Rate	Interrater Correlation	Alpha Reliability
Prose topics			
All papers	4.2	.73	.84
Topic A	3.5	.75	.86
Topic B	4.9	.71	.83
Chart/Graph topics			
All papers	7.5	.67	.81
Topic C	11.4	.60	.75
Topic D	3.6	.75	.86
30-minute condition[a]			
Prose topics	4.4	.72	.84
Chart/Graph topics	6.9	.70	.82
45-minute condition[a]			
Prose topics	5.0	.73	.84
Chart/Graph topics	8.0	.66	.80

[a]Data for 30-minute or 45-minute essays for all experimental conditions except the supplemental condition.

batches of essays. Despite these observations, the levels of interrater consistency obtained were sufficient for purposes of addressing the issues under study here.

More pertinent than the absolute levels of these indices were comparisons--particularly that between essays written under 30- versus 45-minute conditions. Comparison of these two conditions showed little difference in discrepancy rates, interrater correlations, or alpha coefficients. Apparently, then, allowing 45 minutes did not produce essays that yielded notably higher (or lower) scoring reliability than did allowing 30 minutes.

Correlational Data, Including Parallel-Form Reliability

Correlations were computed between the 30- and 45-minute conditions for students who wrote one essay under each time condition (i.e., students in the main treatment groups of the study). Then, to provide a basis for comparison, correlations were computed between essays for students who wrote two 30-minute essays and for students who wrote two 45-minute essays, to provide evidence of parallel-form reliability. To the extent that the correlations between 30- and 45-minute conditions exceeded the parallel-form reliabilities, it could be said that the provision of 15 extra minutes affected the relative standing of the students.

The correlational data are shown in Table 3. The correlation for each experimental condition is based on all students in that condition, including intensive English and academic students combined (thus providing a sufficient range of scores). Each mean correlation was computed by taking the weighted average of transformed (z) scores and retransforming to an r statistic.

For the prose topics, the average correlation between 30- and 45-minute essays was relatively high and was roughly the same as the average correlation between two 30-minute essays and between two 45-minute essays. Note that differences among the four correlations between scores under the two time limits were likely due to random sample variation, as there was no reason to assume that the order of topics or the order of time allotments should have affected the correlations. The mean correlation of .77 between 30- and 45-minute essays was thus somewhat misleading, because it was partly due to the unusually high correlation of .87 for one of the four groups. Still, the median correlation between 30- and 45-minute prose essays for the remaining three groups was .73, which compares favorably with the average correlation of .75 for the groups who wrote two essays under the same time limit. For the chart/graph topics the mean correlation of .69 between 30- and 45-minute essays approached the average correlation of .74 for groups who wrote essays on two chart/graph topics under the same time limit.

In general, then, the correlation between 30- and 45-minute essays was not substantially lower than the correlation between two 30-minute or two 45-minute essays. The latter correlation serves as an index of parallel-form reliability and, in effect, represents the highest level one might

Table 3

Correlation Between Essays for Each Condition

Condition	Prose Topics		Chart/Graph Topics	
	Topic A then B	Topic B then A	Topic C then D	Topic D then C
30-min. essay then 45-min. essay	.73	.87	.68	.69
45-min. essay then 30-min. essay	.71	.74	.64	.73
Mean correlation between 30- and 45-min essays[a]	.77		.69	
Two 30-min. essays	.76	.72	.69	.71
Two 45-min. essays	.78	.74	.76	.78
Mean correlation between essays written under same time limit[a]	.75		.74	

[a]Average of the four correlations immediately above it.

expect for the relation between 30- and 45-minute essays (although it is not an actual limit). Compared with this level, the relation between 30-and 45-minute essays appeared to be about as high as could be expected. This finding suggests that the students rank-ordered in approximately the same way under 30- and 45-minute conditions. To illustrate, the order of scores received by a given pair of students generally tended to be the same for the essay written under the 30-minute condition as for the essay written under the 45-minute condition (whatever the effects of extra time on the mean scores). Although this rule did not apply to every possible pair of students, it applied roughly as often as it did when the two essays were both written under the same time limit. This result is particularly important, because it shows that neither time condition was superior to the other with respect to measuring the students' writing ability in relation to each other.

The parallel-form reliabilities were obtained mainly to provide a basis for comparison with the correlations between time conditions, as discussed above. Nevertheless, they provide important data in themselves for comparison between time conditions in reliabilities. For the prose topics--the topics of primary interest in the study--the parallel-form reliability obtained with the two 30-minute essays averaged .74, whereas that obtained with the two 45-minute essays averaged .76. These figures were nearly identical, suggesting that extending the time limit from 30 to 45 minutes had essentially no effect on the test's parallel-form reliability. The comparable figures for the chart/graph topics were .70 and .77. Although somewhat more discrepant than those for the prose topics (for reasons that would need further research to explain), these figures were still relatively comparable. In general, then, the data tend to support the view--particularly for the prose topics--that the consistency of measurement provided by the TWE is approximately the same under 30-minute and 45-minute time limits.

Performance Effects for Main Treatment Groups

Mean scores on the two essays for the main treatment groups are presented in Table 4; data for the prose topics are presented in Table 4a and data for the chart/graph topics, in Table 4b. It is notable that the mean score for the 45-minute essay was higher than the mean score for the 30-minute essay for 15 of the 16 subgroups in Tables 4a and 4b. That is, the mean second-essay score was generally higher when the essays were presented in order 30 then 45 minutes, whereas the mean first-essay score was generally higher when the essays were presented in order 45 then 30 minutes.

Table 4a

Performance Data for Main Treatment Groups:
Prose Topics

	Topic A then B			Topic B then A		
	N	Mean	SD	N	Mean	SD
30-min. essay then 45-min. essay						
Int. Engl. students						
30-min. essay	20	3.38	.76	18	3.19	.81
45-min. essay	20	3.75	.82	18	3.44	.54
Academic students						
30-min. essay	13	3.92	.89	13	4.58	.84
45-min. essay	13	4.19	.83	13	4.77	.87
45-min. essay then 30-min. essay						
Int. Engl. students						
45-min. essay	29	3.40	.94	26	3.40	.69
30-min. essay	29	3.19	.83	26	3.29	.64
Academic students						
45-min. essay	12	5.00	.77	12	4.96	.96
30-min. essay	12	4.67	.86	12	4.46	.81

Table 4b

Performance Data for Main Treatment Groups:
Chart/Graph Topics

	Topic C then D			Topic D then C		
	N	Mean	SD	N	Mean	SD
30-min. essay then 45-min. essay						
Int. Engl. students						
30-min. essay	20	3.48	.99	15	3.17	.77
45-min. essay	20	3.65	.99	15	3.80	1.07
Academic students						
30-min. essay	13	4.38	.92	12	4.25	1.01
45-min. essay	13	4.73	.95	12	4.50	.95
45-min. essay then 30-min. essay						
Int. Engl. students						
45-min. essay	24	3.65	.83	26	3.46	.81
30-min. essay	24	3.29	.61	26	3.23	1.03
Academic students						
45-min. essay	11	4.32	.90	11	5.05	.88
30-min. essay	11	4.50	.50	11	4.23	1.23

Main analyses of variance. Analyses of variance were conducted separately for the prose and chart/graph topics. The dependent variable in each case was the difference between the first and second essays[2], and the factors were (a) time order (i.e., 30 then 45 minutes versus 45 then 30 minutes); (b) topic order (i.e., A then B versus B then A in the prose analysis; C then D versus D then C in the chart/graph analysis); and (c) educational status (intensive English versus academic students). The least-squares method underlying the analyses of variance effectively ensured that each treatment group in an analysis contributed equally in computing the effects; thus, the different orders (i.e., different time orders, different topic orders) balanced each other, as intended, despite group differences in numbers of students. Wherever mean scores for combinations of groups are presented in this report, they have been computed as the simple, unweighted means of the groups involved.

The central question regarding performance effects was whether scores were greater under the 45- than the 30-minute time limit. This question was addressed here via the time-order effect: If scores averaged higher on the first of the two essays when it was written under the 45-minute time limit, but lower when it was written under the 30-minute time limit, this would be reflected in a significant effect of time order (on the first- minus second-essay difference score). For both the prose and chart/graph analyses, the time-order effect was highly significant--prose: F (1,135) = 20.38, p < .001; chart/graph: F (1,124) = 20.23, p < 001. Thus, adding an extra 15 minutes reliably increased the students' scores.

In neither analysis was there a significant interaction between time order and educational status. This shows that the difference between 45- and 30-minute conditions was not markedly greater or less for the academic students than for the intensive English students. Other effects in the analysis of prose topics were nonsignificant. In the chart/graph analysis, two other effects were significant beyond the .05 level: (a) time order x essay order, the difference between 45- and 30-minute conditions being more pronounced for students receiving Topic D first than C first, F (1,124) = 4.50, p < .05, and (b) essay order x educational status, the scores on Topic C being higher than those on Topic D for intensive English students but the

---

[2]It would also have been possible to include first versus second essay as a within-subject factor in a repeated-measures analysis. However, the effects of this factor in itself were not of particular interest here; of greatest relevance to the study were effects involving the difference between essays. (Note that interactions of first versus second essay with the between-subject factors are mathematically equivalent to effects of the between-subject factors on the difference score.) It might be noted that the mean first and second essay scores were nearly identical. For the main experimental groups (i.e., those involving one 30-minute essay and one 45-minute essay), the mean difference between first and second essays was only .01 points for the prose topics, and -.02 points for the chart/graph topics. For the groups given two essays under the same time limit (30 minutes or 45 minutes), the mean difference between first and second essays was only .06 points for the prose topics and -.04 points for the chart/graph topics.

reverse for academic students, F (1,124) = 8.26, $p < .01$. No obvious explanation for these last two effects is apparent. The effects do not, however, alter the conclusion regarding the central issue--that allowing 45 minutes produced significantly higher scores than did allowing 30 minutes.

Along with determining statistical significance, it is important to consider the magnitude of the differences between the 30- and 45-minute conditions to get an idea of the practical significance of the effects observed. For the prose topics, average scores for the 30- and 45-minute conditions, respectively, were 3.84 and 4.12, for a difference of .28 points on the 6-point TWE scale. For the chart/graph topics, average scores for the 30- and 45-minute conditions, respectively, were 3.82 and 4.25, for a difference of .33 points. Given that the average standard deviations were .80 and .90 for prose and chart/graph topics, respectively, the increase due to adding 15 minutes was equal to .35 standard deviations for the prose topics and .37 standard deviations for the chart/graph topics. These increases exceeded the 1/4 standard deviation that is often regarded as a minimum for defining a meaningful difference, although they were of relatively modest magnitude.

Analysis by level of proficiency. Although evidence of the role of English proficiency was provided above in effects involving educational status (i.e., intensive English versus academic students), it is also of value to look for differences in effects due to writing proficiency. Toward that end, analyses of variance were conducted that were similar to those described above (separately for prose and chart/graph topics), but with level of performance in the present situation rather than educational status as the third factor. In this case, performance level was defined as obtaining an average essay score of less than 3.75 (the median score, both on the prose topics and on the chart/graph topics) versus a score equal to or greater than 3.75. The numbers of students in these two groups were 70 and 73 in the analysis of prose topics and 58 and 74 in the analysis of chart/graph topics.[3]

As expected, the time-order effect was significant in both analyses, again reflecting the difference between 45- and 30-minute conditions--prose: F (1,135) = 18.62, $p < .001$; chart/graph: F (1,124) = 18.43, $p < .001$. However, the interaction between time order and performance level was not significant in either analysis ($F < 1$), showing that the score difference in favor of the 45-minute condition was no more (or less) pronounced for high-scoring than for low-scoring students. Thus, there was no indication that the effects of extra time were related to the students' writing proficiency. For prose topics, the score differences were .25 and .27 for low- and high-scoring students, respectively; for chart/graph topics, the comparable

[3]There was a substantial degree of overlap between the educational status factor and the performance level factor. For the prose topics, only 34% of intensive English students but 82% of academic students scored at or above the median performance level of 3.75; for the chart/graph topics, the figures were 40% and 85%, respectively.

figures were .28 and .35. No other effect in either analysis was significant.

Other performance analyses. In the questionnaire, which will be discussed at length below, the students were asked whether 30 minutes was enough time to write an essay. Students were classified according to whether they felt it was (a) "much less time" or "less time" than they needed versus (b) "the right amount of time," "more time," or "much more time" than they needed. (The numbers of students in these two groups were 79 and 64 for the prose analysis, and 70 and 60 for the chart/graph analysis.) Separate analyses of variance for prose and chart/graph topics were conducted that were analogous to those described above but with questionnaire response in place of educational status as the third factor. The analysis showed, as expected, a significant time-order effect, again reflecting the higher mean essay score for the 45- than the 30-minute condition--prose: F (1,135) = 17.51, $p < .001$; chart/graph: F (1,122) = 19.30, $p < .001$. No other effects in the analyses approached significance. Most notably, there was no hint of an interaction between time order and response group ($F < 1$ in each analysis). Thus, the effect of extra time was not substantially less pronounced for students who felt 30 minutes was adequate than for students who did not. The mean increase in scores due to extra time was .25 for the former students and .26 for the latter in the prose analysis, and .30 versus .35 in the chart/graph analysis.

The data were also analyzed for a particular target group of students--those with average essay scores between 3.5 and 5.0, the range within which decision thresholds set by TWE score users often fall. Analyses of variance were conducted, separately for prose and chart/graph topics, in which the dependent variable was the difference between first and second essays and the factors were (a) time order, and (b) topic order. The only significant effect in each analysis was the time-order effect, again showing a higher score for students given 45 minutes than 30 minutes; prose: F (1,73) = 6.75, $p < .05$; chart/graph: F (1,68) = 6.98, $p < .05$. The mean difference in scores due to extra time was .24 for the prose topics and .27 for the chart/graph topics.

It was desirable to assess the degree to which the nature of the students might have influenced the results. A variable believed to play a role in this regard was the students' major-field area, as essay writing capability may differ according to students' academic interests and experiences. Major-field information was available only for academic students, so the sample for analysis was limited in this case. These students were divided into two groups: (a) humanities and social science majors and (b) biological and physical science majors. The numbers of students in these groups, respectively, were 22 and 27 for the prose topics, and 24 and 23 for the chart/graph topics. Because of the limited sample, the prose and chart/graph topics were combined into a single analysis. The dependent variable was the difference between first and second essay. Factors were (a) time order, (b) major-field area (humanities/social sciences versus biological/physical sciences, (c) topic type (prose versus chart/graph), and (d) topic order nested within topic type. Although the Ns per cell ranged widely (from 3 to 10), these Ns were within acceptable

limits for conducting such an analysis according to Bartlett (cf., Winer, 1962), and the data met the conditions for homogeneity of variance by Bartlett's test, X^2 (15, N = 96) = 18.98, $p > .10$.

The analysis, as expected, showed a highly significant effect of time order, F (1,80) = 11.82, $p < .001$, again indicating a higher score for the 45- than the 30-minute condition. The effect of particular interest was the interaction between time order and major-field area. This interaction approached significance, F (1,80) = 3.36, $p = .07$, suggesting at least a tendency--to be explored in further research--toward a greater benefit of extra time for humanities/social science students than for biological/physical science students. The mean increase in scores produced by extra time was .46 for the former students and .14 for the latter (prose: .48 versus .19; chart/graph: .44 versus .08). The one other significant effect in this analysis was that of topic order within topic type, F (2,80) = 3.42, $p < .05$, as the difference between first and second essays was greater when prose topics were given in order B-A than the reverse, and greater when chart/graph topics were given in order D-C than the reverse. No explanation for this effect is readily apparent.

Analyses Involving the Supplemental Condition

In the supplemental condition, students were given 45 minutes for their first essay, the first 15 minutes of which were to be used for planning. The question of interest was whether scores on this special 45-minute essay would differ from scores on a regular 45-minute essay, as obtained by students in the main treatment group labeled "45-minute essay then 30-minute essay." Because this comparison involved different groups of students (drawn from different institutions), the second, 30-minute essay was used as a control factor. Analyses of covariance were performed on the data in these two conditions, separately for prose and chart/graph topics. The dependent variable was the score on the first essay; independent factors were instruction type (i.e., instructions to plan first versus no special instructions), topic order, and educational status; and the covariate was the score on the second (30-minute) essay.[4]

The principal effect to be examined was that of instruction type. In neither the prose nor the chart/graph analysis was this effect significant. Thus, the special and regular 45-minute conditions did not differ markedly when controlling for performance on the standard 30-minute test; being required to plan for 15 minutes then having 30 minutes to write did not result in higher or lower scores than did having the full 45 minutes to

[4]The second essay was comparable for both conditions in that it was a standard 30-minute essay. Because this essay came after two different types of first essay, it was not strictly identical for the two groups. This situation did not result in violation of the assumption of homogeneity of regression slopes, and to that extent the use of analysis of covariance was regarded as appropriate here. Ideally, however, the covariate should be an independent measure, obtained under identical conditions for both groups.

write. The only significant effects were (a) educational status--prose topics: F (1, 134) = 19.62, $p < .001$; chart/graph topics: F (1, 116) = 8.08, $p < .01$--and (b) the interaction of topic order and educational status for chart/graph topics, F (1, 116) = 7.66, $p < .01$. These last two effects indicated that academic students outperformed intensive English students on the first essay (the difference being attenuated but not eliminated with covariation of the second essay score), and, for the chart/graph topic type, this effect was due primarily to students who received the topics in order D then C.

Although no difference was observed between the special and regular 45-minute conditions here, this result should be regarded as highly tentative. Comparison across different groups of students, where random assignment to groups is impossible, provides a weak basis on which to draw conclusions about differences, even when a covariate is employed. (By contrast, in the comparisons of regular 30- and 45-minute conditions the students served as their own "controls," thus ensuring relatively sensitive tests.) Conclusions about allocation of time for planning must await research that employs a more effective experimental design.

Questionnaire Data

Table 5 presents questionnaire data from the main treatment groups for questions concerning the adequacy of the time allowed. The data are presented separately for intensive English and academic students but combined across time orders and topic orders. Because factors such as time order and topic order should not play a major role in the students' questionnaire responses, it was deemed appropriate to combine the data in this manner. The Ns shown in the table are the total numbers in the different groups; at least 97% of students in each group responded, and the figures in the table are the percentages of students giving each response.

Responses to Questions 1 and 2 in Table 5 show that the students generally believed 45 minutes to be a more adequate amount of time than 30 minutes, particularly for the prose topics. For the prose topics, a majority--54% of intensive English students and 58% of academic students--felt they had less (or much less) time than needed when given 30 minutes; in contrast, only 19% of intensive English students and 4% of academic students felt similarly about 45 minutes.

For the chart/graph topics, the pattern of responses for intensive English students was similar to that for the prose topics: 67% felt they had less time than they needed when given 30 minutes, but only 19% felt similarly about 45 minutes. For academic students given the chart/graph topics, on the other hand, the comparable figures were only 29% and 6%, as these students generally found 30 minutes to be adequate, and 45 minutes more than adequate, to write about the chart/graph topics.

With the exception of the chart/graph topic for academic students, then, 30 minutes was generally perceived by the students to be insufficient time, but 45 minutes sufficient, for writing essays on these topics.

Table 5

Responses to Questions Regarding Time Allotment for Main Treatment Groups: Percentages of Students Responding to the Different Options

	Prose Topics		Chart/Graph Topics	
	Int. Engl. Students (N = 93)	Acad. Students (N = 50)	Int. Engl. Students (N = 85)	Acad. Students (N = 47)
1. When I was given 30 minutes I felt I had:				
much less time than I needed	13	8	13	6
less time than I needed	41	50	54	23
the right amount of time	35	36	28	55
more time than I needed	10	6	5	11
much more time than I needed	1	0	0	4
2. When I was given 45 minutes I felt I had:				
much less time than I needed	5	2	2	2
less time than I needed	14	2	17	4
the right amount of time	52	58	64	32
more time than I needed	24	36	14	45
much more time than I needed	5	2	2	17

In the questionnaire, the students were also told to "Imagine that, as part of the TOEFL, you are asked to write an essay like the ones you wrote today," and they were asked the following two questions: (a) "To write an essay as part of the TOEFL, 30 minutes would be," followed by the five options shown in Table 4, and (b) "To write an essay as part of the TOEFL, 45 minutes would be," followed by the same five options. These questions were included as a check on the possibility that the students' perceived time requirements in an operational testing situation might be quite different from those in the present experimental situation. This was not the case, however, as the results closely matched those shown in Table 4. When percentages were computed that paralleled those in Table 4, all but 6 of the 40 numbers were within 4 percentage points of their counterparts in Table 4, with no consistent pattern to the differences.

The apparent differences between intensive English and academic students are of particular interest, because they bear on the hypothesis that students' proficiency relates to their perceptions about the adequacy of the time allowed (where educational status serves as a rough external index of English proficiency). Statistics were computed to compare the responses of intensive English and academic students. For each of the two questions in Table 5, for each topic type, a chi-square statistic compared these two groups with respect to (a) the number of students responding to one of the first two options ("much less time than I needed" or "less time...") versus (b) the number responding to the other three options. The group difference was not significant for Question 1, regarding 30 minutes, for the prose topics. The difference was, however, significant (a) for Question 1 for the chart/graph topics, X^2 (1, N = 130) = 17.14, p <. .001, (b) for Question 2 (regarding 45 minutes) for the prose topics, X^2 (1, N = 143) = 6.37, $p < .05$, and (c) for Question 2 for the chart/graph topics, X^2 (1, N = 130) = 4.00, $p < .05$.

Parallel analyses were conducted with low versus high mean essay score in place of educational status as an index of proficiency. Again, the chi-square statistic was significant for the last three of the four situations mentioned above (with chi-square values ranging from 5.92 to 8.50), but not for the first. The percentages of low- versus high-scoring students reporting 30 minutes to be insufficient were, for the prose topics, 57% versus 52%; for the chart/graph topics, 66% versus 44%. The comparable percentages reporting 45 minutes to be insufficient were, for the prose topics, 22% and 7%; for the chart/graph topics, 25% and 7%. Thus, while the overall picture showed that both low- and high-proficiency students regarded 45 minutes as sufficient for both topics and 30 minutes as insufficient for the prose topic, there was still some group variation, with lower proficiency students somewhat less satisfied than higher proficiency students with the time limits.

The questionnaire data for the main treatment groups, discussed above, are the most relevant to comparisons of the 30- and 45-minute conditions, as these groups had first-hand experience with both time conditions. Nevertheless, it is possible that the students in these groups felt obliged to respond differentially to questions about the 30- and 45- minute conditions, because they could readily infer that the purpose of the study

was to compare these two time conditions. For this reason, it was useful also to examine the questionnaire data for students who had only 30-minute or only 45-minute time conditions and, thus, had no reason to suspect that the study involved a comparison between time conditions.

Table 6 presents the responses of students given two 30-minute or two 45-minute essays to a question about the adequacy of the time allowed. In this case the Ns presented in the table are the actual numbers of students responding to the questions listed; for each group the number shown is at least 96% of the number in the total group.

The data in Table 6 closely parallel the data in Table 5. For prose topics, and for chart/graph topics for intensive English students, majorities of students found 30 minutes to be less (or much less) time than needed, whereas majorities of students found 45 minutes to be sufficient. And for those given the chart/graph topics, a majority of academic students, but not intensive English students, regarded 30 minutes as sufficient. A difference from the results in Table 5 is that relatively few academic students here reported 45 minutes to be too much time for the chart/graph topics. Also, for the prose topics here, the intensive English students showed a greater tendency than academic students to regard 30 minutes as insufficient.

Additional questions posed to students in the main treatment groups addressed the issue of how the students used the additional 15 minutes provided in the 45-minute condition. These questions are presented in Table 7. (The percentages of students responding to these questions ranged from 91% to 98%.) Perhaps the most important aspect of these data is that the additional 15 minutes was used for all three purposes. Although some differences were observed, examination of all four questions in combination suggested no strong, consistent tendency toward use of the extra time for one purpose more than the others. Thus, the beneficial effects of extra time on performance could have been attributable to a combination of these factors.

Chi-square analyses compared the responses of intensive English and academic students to Questions 1, 2 and 3 in Table 7. For the prose topics, these groups did not differ significantly in percentage of "yes" responses to any of these questions. For the chart/graph topics, the two groups differed only in response to Question 3, regarding editing after writing, X^2 (1, N = 128) = 8.57, $p < .01$. Comparable analyses with low versus high mean essay score as an index of proficiency also revealed no significant differences except on Question 3 for the chart/graph topics: 36% of low scorers but 64% of high scorers used part of the extra time for editing with the chart/graph topics. Thus, for the prose topics, proficiency was unrelated to the number of students who reported using the additional time to plan, write, or edit. For the chart/graph topics, the reported incidence of editing, but not planning or writing, during this extra time was greater for high- than low-proficiency students. This latter result undoubtedly

Table 6

Responses to Questions Regarding Time Allotment for Students Who Wrote Two 30-minute Essays or Two 45-minute Essays: Percentages of Students Responding to the Different Options

	Prose Topics		Chart/Graph Topics	
	Int. Engl. Students	Acad. Students	Int. Engl. Students	Acad. Students
Two 30-minute Essays	N = 49	N = 49	N = 44	N = 48
For the essays I wrote today, I felt I had:				
much less time than I needed	12	10	16	10
less time than I needed	59	41	48	19
the right amount of time	29	39	27	48
more time than I needed	0	10	9	19
much more time than I needed	0	0	0	4
Two 45-minute Essays	N = 71	N = 48	N = 67	N = 47
For the essays I wrote today, I felt I had:				
much less time than I needed	4	2	3	2
less time than I needed	23	25	28	4
the right amount of time	52	56	49	70
more time than I needed	15	17	16	15
much more time than I needed	6	0	3	9

Table 7

Responses to Questions Regarding Use of Additional Time for Main Treatment Groups

	Prose Topics		Chart/Graph Topics	
	Int. Engl. Students (N = 93)	Acad. Students (N = 50)	Int. Engl. Students (N = 85)	Acad. Students (N = 47)
Recall what you did today when you were given 45 minutes, and compare it with what you did when you were given 30 minutes.				
Data for questions 1 through 3 are percentages of "yes" responses.				
1. Did you use part of the extra 15 minutes to plan and make notes before writing?	48	48	52	52
2. Did you use part of the extra 15 minutes to write a longer essay?	51	54	46	41
3. Did you use part of the extra 15 minutes to edit or correct your essay after writing a draft?	52	53	43	67
4. Which one of the following things did you spend the most time doing during the extra 15 minutes? (CHECK ONLY ONE)				
Data are percentages responding to the different options.[a]				
planning and making notes before writing	38	26	41	33
writing a longer essay	33	45	30	28
editing and correcting my essay after writing a draft	29	30	29	40

[a]Due to rounding, the total percentage does not equal 100 in all cases.

relates to the observation that many of the higher proficiency students found 45 minutes to be more than adequate, thus leaving them with much time available for continued work after completing a draft.

For the supplemental condition, in which students were required to spend the first 15 minutes planning, the students were asked whether the planning period was (a) very useful, (b) somewhat useful, or (c) not very useful. Percentages responding to each category were, for the prose topics, 63, 30, and 8; for the chart/graph topics, 57, 34, and 0. Students also responded to the statement "The 15 minutes allowed for planning was (a) much less time than I needed, (b) less time than I needed, (c) the right amount of time, (d) more time than I needed, or (e) much more time than I needed." Percentages responding to each category were, for the prose topics, 2, 8, 59, 27, and 5; for the chart/graph topics, 0, 9, 51, 34, and 6. Thus, the students found the planning period to be very useful, and the majority reported that 15 minutes was the right amount of time, although among those giving other responses, there was a tendency to regard 15 minutes as too much time rather than too little time.

Discussion

Correlations and Reliability

In assessing the psychometric integrity of the TWE under 30- versus 45-minute time limits, the first question of interest concerns the correlation between scores under these two conditions. It was reasoned that, if the correlation between 30- and 45-minute test scores were relatively low, the ranking of students on the test could be said to have been affected by the provision of extra time. Such was not the case, however. The correlation in performance under 30- versus 45-minute time limits was nearly as great as the correlation between scores on two 30-minute or two 45-minute tests. Because the latter two cases serve as indices of parallel-form reliability, correction for unreliability would yield a correlation between 30- and 45-minute conditions that is near unity. Apparently, while provision of an extra 15 minutes increased the students' average scores, it did so relatively uniformly across the spectrum of ability and did not markedly affect the relative standing of students on the test. In this key respect, then, the time extension did not alter the basic character of the test. The test administered with a 30-minute time limit was essentially comparable to a test with a 45-minute time limit with respect to measuring the writing ability of the students in relation to each other.

It is also of interest to look at data on test reliability under each time condition. The parallel-form reliability for the 30-minute condition (i.e., the correlation between the two 30-minute tests) was comparable to the parallel-form reliability for the 45-minute condition--particularly for the prose topics, which were the topics of primary interest in the study. (For the chart/graph topics, the effect of time limits on reliability was slightly greater but was still not substantial.) Thus, the consistency of measurement provided by the TWE appears to have been approximately the same for the 30- as for the 45-minute time condition. Interrater reliabilities were also similar for the two time conditions, indicating that scoring consistency was comparable for essays written under both conditions.

Mean Performance Effects

Another issue of interest concerns the effects of the additional 15 minutes on the mean essay scores. On average, provision of the additional 15 minutes raised scores by .28 points on the 6-point TWE scale for the prose topics and .33 points for the chart/graph topics. These increases were equal to approximately 1/3 of a standard deviation, a difference that is generally regarded as practically meaningful, albeit relatively modest. The positive effect of extra time is consistent with the results reported in certain studies cited above.

Because a majority of students reported that 30 minutes was less time than needed, the effect of the extra 15 minutes on mean scores might seem to have been due to making these students feel less rushed. That this interpretation is incorrect, however, is shown in the finding that extra

time produced roughly the same increase in scores for students who reported that 30 minutes was enough time as for students who did not. Thus, whatever role was played by the provision of extra time, it apparently was not to reduce or eliminate any inequity that may have existed due to some students having insufficient time.

Extra time can lead to increased scores in any of several ways. It can give the student a greater chance to prepare before writing. It can provide an oppo tunity to write a longer essay, and to develop an argument more fully. And it can permit more time for editing after writing. Responses to the questionnaire suggested that, across students, each of these three functions came into play, with no clear indication that one of them played a consistently greater role than the others.

The role of student proficiency. It was thought that extra time might differentially affect students of low versus high proficiency. The present data, however, showed little relationship between proficiency and effects of extra time on performance, where proficiency was defined either as the average score on the two essays written here or by enrollment in intensive English versus academic programs. Low-proficiency students did report somewhat less satisfaction than did high-proficiency students with the time allowed in certain cases. Even then, however, the pattern of results was generally comparable for the two groups: roughly equal majorities of both low- and high-proficiency students reported 30 minutes to be less time than needed for the prose topics, and large majorities of both groups found 45 minutes to be sufficient.[5] Overall, then, the most notable aspect of the comparison between proficiency levels seems to be that, in most respects, low- and high-proficiency students reacted similarly to the provision of additional time--both in mean scores and in response to questions about use of the extra time and the adequacy of the time allowed.

Practical Implications

A key question underlying this study was whether it is essential to extend the time limit on the TWE beyond 30 minutes to provide adequate measurement of students' writing ability. It is useful to summarize those aspects of the present results that bear on this question.

Of particular importance were the results based on the correlational data. These results suggested that the addition of 15 minutes to the test did not substantially alter the relative standing of the students. The 30-minute condition was essentially comparable to the 45-minute condition with respect to measuring the writing ability of students in relation to each other. Thus, in regard to measuring individual differences in writing ability, the psychometric character of the TWE under a 30-minute time limit

---

[5]The main difference was that, for the chart/graph topics written under a 30-minute time limit, a substantial majority of low-proficiency students reported the time to be less than needed, whereas a large majority of high-proficiency students found the time to be adequate.

appeared to be comparable to that under the more extended, 45-minute limit.

The two time conditions were also reasonably comparable in reliability. Parallel-form reliability was approximately the same under the 30-minute condition as under the 45-minute condition (particularly for the key, prose topic type), showing that the consistency of measurement provided by the test was approximately the same for these two conditions. Also, similar interrater reliabilities showed that scoring consistency was comparable for essays written under the 30- and 45-minute time limits. With respect to reliability, then, the psychometric character of the test also appeared to be little affected by the addition of 15 minutes.

Adding 15 minutes to the test did have the effect of increasing the mean scores. The means were increased by a relatively modest amount--about 1/4 to 1/3 scale point, on average, or 1/3 of a standard deviation. Apparently, the additional 15 minutes provided a greater opportunity to plan, write, and/or edit, and, judging from the questionnaire responses, the students (as a group) did all three. In short, the extra time provided an opportunity for fuller development of an essay.

Whether this result shows the current 30-minute time limit to be inadequate for effective measurement of writing ability, however, is another question. If the effect of the time extension had been especially pronounced for low-proficiency students, it might be argued that those students were differentially disadvantaged and needed more than 30 minutes to demonstrate their writing capability. However, the finding that the additional 15 minutes did not benefit those students more than the high-proficiency students suggests that the extra time may simply have afforded students at all proficiency levels an opportunity for fuller essay development. Similar reasoning might apply to the finding that extra time did not provide any greater benefit for students who reported 30 minutes to be less time than needed than for the other students. This result suggests that the effect of extra time was not so much to reduce the tendency for certain students to be "rushed" as to allow all students an opportunity for further development of their essays. Implications of the results involving mean performance are discussed further below.

Also to be considered are the student questionnaire data, which showed that the majority of students reported 30 minutes to be less time than needed (and 45 minutes to be enough time). To the extent that students' perceptions are regarded as useful evidence in this regard, the face validity of the test appeared to be greater with the more liberal time limit. Unfortunately, the present study was unable to show whether the students who responded "less time than needed" meant that 30 minutes was (a) inadequate time to demonstrate their writing ability satisfactorily or (b) simply less time than desired to present a fully developed argument. The former would have greater consequences than the latter for face validity. Thus, as discussed below, it is an important issue for further research to determine which of these alternatives more accurately characterizes the students' perceptions.

A special consideration regarding scoring of TWE essays. If the TWE were norm referenced, as are most standardized tests, data regarding mean performance would be relatively insignificant in relation to the correlational data for determining whether extra time affects the psychometric integrity of the test. The fact that the relative standing of the students was generally uninfluenced by extra time, as shown in the correlations, would suggest that the essential measurement properties of the test were not affected by the change in time limits. The increase in mean scores with extra time would be comparatively unimportant from a measurement perspective, given that the performance of students in relation to each other remained relatively unchanged.

Because the TWE scale is criterion referenced, however, an effect of extra time on the mean score has potentially greater significance. Each level of the 6-point TWE scale is defined by a given set of descriptors regarding aspects of writing competence. (See Test of Written English Scoring Guide in Appendix B.) For example, a score of 4.0 is given to an essay that is minimally adequate in key respects; scores above 4.0 reflect more than minimally adequate competence, and scores below 4.0, less than minimally adequate competence. If score users tend to rely on a fixed cutoff score in making decisions about students, and if students are more likely to achieve that score with a 45- than a 30-minute test, ostensibly it might seem that providing the additional time is important for achieving an accurate evaluation of the examinees' writing ability.

However, in making this interpretation, one must take into account the nature of the scoring procedure. When the essay readers evaluate students' essays in a standard TWE reading, it is with the expectation that the students have been given 30 minutes. (Indeed, the reading supervisor explicitly stresses, in discussing use of the scale descriptors, that students were working within a 30-minute time restriction.) It is reasonable to hypothesize that readers allow their interpretation of the scale descriptors to be influenced by their expectations about what students can accomplish within a 30-minute time constraint. In short, although the scale descriptors do reflect aspects of performance that are nonrelative to a population's score distribution, the descriptors may still be interpreted and applied relative to the time limits under which the test was administered.

Issues for Further Research

An issue for further research is whether changes in essay readers' expectations about the time constraints indeed affect the way in which they apply the TWE scale descriptors. In the present study, as in any standard TWE administration, scoring was done by expert readers whose prior experience and training involved rating 30-minute essays. If the testing conditions were changed to allow more time for all examinees, such as 45 minutes, it is possible that essay readers would adjust their expectations accordingly. For example, readers might well modify what they expect from an essay when they apply the descriptors "adequately organized and developed," "uses some details to support a thesis or illustrate an idea,"

and so forth (cf., Test of Written English Scoring Guide in Appendix B). If so, 45-minute essays written and scored under the revised testing conditions might well receive somewhat lower scores than 45-minute essays that are evaluated in the manner emplcyed in this study--that is, rated simultaneously with 30-minute essays, in a time-blind scoring procedure, by readers accustomed to evaluating 30-minute essays.

To resolve this issue satisfactorily, an extensive research effort would likely be required. A means of addressing the issue would be to create conditions under which some readers would receive training and a considerable amount of experience with 45-minute essays before scoring papers that were written under the 45-minute condition, and other readers would be given training and experience with 30-minute essays before scoring papers written under the 30-minute condition. (Of course, steps would need to be taken to ensure the comparability of the two reader groups.) The feasibility of such a complex research project remains to be determined. Nevertheless, without resolution of this issue, effects of time limits on the mean essay scores, such as those observed in the present study, should not serve as a major basis for drawing conclusions about the adequacy of the 30-minute time limit.

Another issue to be resolved concerns students' perceptions about the time limits. Although a majority of students in this study responded that 30 minutes was less time than they needed, this response can be interpreted in different ways, as discussed above. Additional research is needed to determine whether students actually feel so rushed with a 30-minute time limit that they cannot demonstrate their writing capabilities satisfactorily, or whether they simply feel that 30 minutes is less time than they would prefer to have in order to present as fully developed an essay as they would like. To resolve this issue requires more than a simple questionnaire question. Ideally, it would involve observation and interviewing of students in connection with the essay-writing experience, using a variety of questions, to get a more accurate picture of whether the students perceive 30 minutes to be truly inadequate.

Even if the research suggested above were to yield results pointing to the value of extending the TWE time limit, it would remain to be determined how much additional time should be provided. The decision to compare the standard 30-minute condition with a 45-minute condition in the present study was arbitrary, designed to provide initial evidence regarding effects of the time limits on the TWE. If it appeared desirable to extend the time limit beyond 30 minutes, it is possible that an extension of just 5 minutes, or 7 or 10 minutes, rather than 15 minutes would be sufficient. Therefore, comparison of performance in, and students' reactions to, several different time conditions would be needed before a decision could be made about a suitable amount of time to allow.

Finally, a potentially important research issue is whether effects of time limits on the TWE vary for students of different backgrounds. In the present study, a nearly significant relation was observed between the student's major-field area and the effects of time allowed, a result that merits follow-up investigation. It would also be of value to examine the

relationships of such student characteristics as native language or region to the effects of the time limits. Although the students in the present study were not sampled in large enough numbers to permit proper tests in this regard, more systematic and comprehensive sampling by key language groups or regions would allow examination of these relationships. Such research would be valuable in determining whether students of different backgrounds would be differentially advantaged by the provision of time beyond the 30 minutes currently allowed.

References

Biola, H. R. (1982). Time limits and topic assignments for essay tests. Research in the Teaching of English, 16, 97-98.

Caudery, T. (1990). The validity of timed essay tests in the assessment of writing skills. ELT Journal, 44, 122-131.

Collins, A., & Gentner, D. (1980). A framework for a cognitive theory of writing. In L. W. Gregg & E. R. Steinberg (Eds.), Cognitive process in writing. Hillsdale, NJ: Erlbaum.

Educational Testing Service (1989). TOEFL Test of Written English Guide. Princeton, NJ: Author.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In Kroll (Ed.), Second language writing: Research insights for the classroom. Cambridge, England: Cambridge University Press.

Livingston, S. A. (1987, April). The effects of time limits on the quality of student-written essays. Paper presented at the meeting of the American Educational Research Association, Washington, DC.

McLaughlin, B. (1987). Theories of second-language learning. London: Edward Arnold.

Mitchell, K. J., & Anderson, J. A. (1987, April). Estimation of interrater and parallel forms reliability for the MCAT essay. Paper presented at the meeting of the American Educational Research Association, Washington, DC.

Ruth, L., & Murphy, S. (1988). Designing writing tasks for the assessment of writing. Norwood, NJ: Ablex.

Stansfield, C. W., & Ross, J. (1988). A long-term research agenda for the Test of Written English. Language Testing, 5, 160-186.

Winer, B. J. (1962). Statistical principles in experimental design. New York: McGraw-Hill.

Younkin, W. F. (1986). Speededness as a source of test bias for non-native English speakers on the College Level Academic Skills Test. (Doctoral dissertation, University of Miami). Dissertation Abstracts International, 47/11-A, 4072.

APPENDIX A

Topics Used in Study

Prose Topic A

A company has announced that it wishes to build a large factory near your community. Discuss the advantages and disadvantages of this new influence on your community. Do you support or oppose the factory? Explain your position.

Prose Topic B

Inventions such as eyeglasses and the sewing machine have had an important effect on our lives. Choose another invention that you think is important. Give specific reasons for your choice.

Chart/Graph Topic C

COMPARATIVE RATINGS OF FOUR BRANDS OF BREAD

	Quality of Taste	Nutritional Rating	Cost per Pound
Brand A	★★★★	★★★★	$0.86
Brand B	★★★	★★★	0.79
Brand C	★★	★★★★	0.41
Brand D	★★★★★	★★	0.40

Excellent	Very Good	Good	Fair	Poor
★★★★★	★★★★	★★★	★★	★

You have been asked to recommend a particular brand of bread for use in elementary schools in a small city. Which of the four brands of bread described in the chart would you recommend? Support your choice with information from the chart above.

Chart/Graph Topic D

CHARACTERISTICS OF TWO CAREERS

	Career A	Career B
Travel Opportunities	***	**
Job Security	*	****
Opportunities for Promotion	***	*
Salary	****	**
Job Satisfaction	**	****

Excellent	Good	Fair	Poor
****	***	**	*

The chart indicates the characteristics of two types of careers. Which one would you choose? Give reasons to support your answer.

APPENDIX B

Test of Written English Scoring Guide

TEST OF WRITTEN ENGLISH (TWE)
SCORING GUIDE

REVISED 2/90

Readers will assign scores based on the following scoring guide. Though examinees are asked to write on a specific topic, parts of the topic may be treated by implication. Readers should focus on what the examinee does well.

Scores

6 **Demonstrates clear competence in writing on both the rhetorical and syntactic levels, though it may have occasional errors.**

A paper in this category
- -effectively addresses the writing task
- -is well organized and well developed
- -uses clearly appropriate details to support a thesis or illustrate ideas
- -displays consistent facility in the use of language
- -demonstrates syntactic variety and appropriate word choice

5 **Demonstrates competence in writing on both the rhetorical and syntactic levels, though it will probably have occasional errors.**

A paper in this category
- -may address some parts of the task more effectively than others
- -is generally well organized and developed
- -uses details to support a thesis or illustrate an idea
- -displays facility in the use of language
- -demonstrates some syntactic variety and range of vocabulary

4 **Demonstrates minimal competence in writing on both the rhetorical and syntactic levels.**

A paper in this category
- -addresses the writing topic adequately but may slight parts of the task
- -is adequately organized and developed
- -uses some details to support a thesis or illustrate an idea
- -demonstrates adequate but possibly inconsistent facility with syntax and usage
- -may contain some errors that occasionally obscure meaning

3 **Demonstrates some developing competence in writing, but it remains flawed on either the rhetorical or syntactic level, or both.**

A paper in this category may reveal one or more of the following weaknesses:
- -inadequate organization or development
- -inappropriate or insufficient details to support or illustrate generalizations
- -a noticeably inappropriate choice of words or word forms
- -an accumulation of errors in sentence structure and/or usage

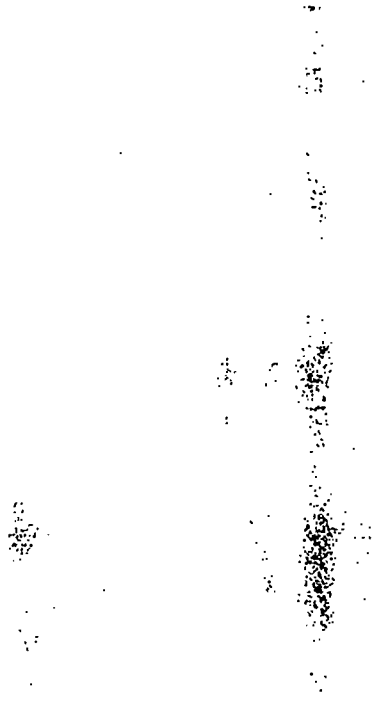
2 **Suggests incompetence in writing.**

A paper in this category is seriously flawed by one or more of the following weaknesses:
- -serious disorganization or underdevelopment
- -little or no detail, or irrelevant specifics
- -serious and frequent errors in sentence structure or usage
- -serious problems with focus

1 **Demonstrates incompetence in writing.**

A paper in this category
- -may be incoherent
- -may be undeveloped
- -may contain severe and persistent writing errors

Papers that reject the assignment or fail to address the question must be given to the Table Leader.
Papers that exhibit absolutely no response at all must also be given to the Table Leader.



Printed on Recycled Paper

57906-01202 • Y62M.5 • 275585 • Printed in U.S.A.